

# Bayesian Robustness to Outliers in Linear Regression

Philippe Gagnon<sup>a,\*</sup>, Alain Desgagné<sup>b</sup>, Mylène Bédard<sup>a</sup>

<sup>a</sup>*Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Québec, Canada, H3C 3J7*

<sup>b</sup>*Département de mathématiques, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal, Québec, Canada, H3C 3P8*

---

## Abstract

Whole robustness is an outstanding property to have for statistical models. It implies that the impact of outliers gradually vanishes as they approach plus or minus infinity, and consequently, the conclusions obtained are consistent with the majority of the observations (the bulk of the data), contrary to nonrobust models. In this paper, we make two contributions. First, we generalise the available whole robustness results, which are for simple linear regression through the origin, to the usual linear regression model. Second, we present a Bayesian approach to robustly identify statistically significant linear dependencies between variables. The strategy to attain whole robustness is simple: replace the traditional normal assumption on the error term by a super heavy-tailed distribution assumption. That apart, analyses are conducted as usual.

**Keywords:** Built-in robustness, Bayesian inference, Pearson's correlation test, super heavy-tailed distributions, maximum likelihood estimation.

**2000 MSC:** 62J05, 62F35

---

## 1. Introduction

The linear regression model is one of the most popular statistical tools. It is commonly used to predict values for a variable (the dependent variable) using auxiliary information (a given set of observations of the explanatory variables), or to quantify the strength of the relationship between the dependent variable and the explanatory variables. The underlying assumption of this model is that the explanatory variables are linearly related to the dependent variable, with an error term that traditionally has a normal distribution. It is now well understood that conflicting sources of information may contaminate the inference when normality is assumed. Indeed, in order to incorporate all the information, an undesirable compromise is made due to the slimness of the tails of the normal: there is a concentration of the posterior on an area that is not supported by any source of information. From a Bayesian perspective, these conflicting sources may represent the prior and outliers; in this paper, we focus on contamination due to the presence of outliers,

---

\*Corresponding author

*Email addresses:* gagnonp@dms.umontreal.ca (Philippe Gagnon), desgagne.alain@uqam.ca (Alain Desgagné), mylene.bedard@umontreal.ca (Mylène Bédard)

and a conflict therefore represents the fact that a group of observations produces a rather different inference than that arising from the bulk of the data and the prior. In particular, in this context of robustness against outliers in linear regression, contaminated inference corresponds to predictions that are not inline with neither the outliers nor the nonoutliers, and misleading quantification of the strength of the relationship between the dependent variable and the explanatory variables. We believe that the appropriate way to address the problem is to limit the influence of outliers in order to obtain conclusions consistent with the majority of the observations.

[Box and Tiao \(1968\)](#) were the first to propose a Bayesian solution. Their recommendation was the following: assume that the distribution of the error term is a mixture of two normals, with one component for the nonoutliers and the other one, with a larger variance, for the outliers. This approach has been generalised by [West \(1984\)](#) who modelled errors with heavy-tailed distributions constructed as scale mixtures of normals, which includes the Student distribution. More recently, [Peña et al. \(2009\)](#) introduced a different robust Bayesian approach where each observation has a weight decreasing with the distance between this observation and the bulk of the data. They proved that the Kullback-Leibler divergence from the posterior arising from the nonoutliers only to the posterior arising from the sample containing outliers is bounded. This result essentially indicates that these posterior densities can be different, but that their ratio has to be bounded.

The main drawback of the proposed approaches is the lack of strong theoretical results that validate the underlying ideas. In this paper, we aim at filling this gap through theoretical results ensuring whole robustness, meaning that the impact of outliers gradually vanishes as they approach plus or minus infinity. To achieve this, we borrow ideas from the paper of [Desgagné and Gagnon \(2016\)](#), in which whole robustness results are provided for simple linear regression through the origin. This aligns our work with the *theory of conflict resolution in Bayesian statistics*, as described by [O’Hagan and Pericchi \(2012\)](#) in their extensive literature review on that topic. The approach is as simple as those of [Box and Tiao \(1968\)](#) and [West \(1984\)](#): replace the traditional normal assumption by a distribution that accommodates for the presence of outliers. The tails of the distribution however have to be heavier than that of the Student distribution, because this distribution only allows to attain partial robustness (see the results of [Andrade and O’Hagan \(2011\)](#) in a context of location-scale model), meaning that outliers have a significant but limited influence on the inference as they approach plus or minus infinity.

In our paper, we first present the general model (with no specific distribution assumption on the error term) in Sect. 2.1. We then describe the distributions that we use to attain whole robustness in Sect. 2.2. They are super heavy-tailed distributions, and more precisely, log-regularly varying distributions (an example of log-regularly varying distribution is given in Sect. 3). When assuming a super heavy-tailed distribution on the error term, the resulting model is characterised by its built-in robustness that resolves conflicts in a sensitive and automatic way, as stated in our robustness results given in Sect. 2.3. The key result is the convergence of the posterior distribution (arising from the whole sample) towards the posterior arising from the nonoutliers only, when the outliers approach plus or minus infinity. This result corresponds to whole robustness and indicates that users can conduct their analyses as usual, estimating parameters with posterior medians and Bayesian credible intervals for instance.

In Sect. 3, we illustrate the practical implications of the theoretical results presented in Sect. 2.3. In particular, in Sect. 3.1, we show the relevance of our approach in analyses of the relationship

between the dependent variable and explanatory variables (in this case, of the relationship between two stock market indexes). We also explain how our approach leads to a simple procedure to robustly identify statistically significant linear dependencies between variables. In Sect. 3.2, we evaluate the accuracy of the estimates arising from a robust multiple linear regression model, based on our approach. Our robust model is compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models. It is showed that our robust model performs as well as the nonrobust and the partially robust models in absence of outliers, in addition to being completely robust. It indicates that, by only changing the assumption on the error term, we obtain adequate estimates in absence or presence of outliers.

## 2. Resolution of Conflicts in Linear Regression

### 2.1. Model

- (i) Let  $Y_1, \dots, Y_n \in \mathbb{R}$  be  $n$  random variables and  $\mathbf{x}_1 := (1, x_{12}, \dots, x_{1p})^T, \dots, \mathbf{x}_n := (1, x_{n2}, \dots, x_{np})^T \in \mathbb{R}^p$  be  $n$  known vectors, where  $p \in \{1, 2, \dots\}$  and  $n > p + 1$  are assumed to be known. We want to study the relationship between  $\mathbf{Y}_n$  and  $\mathbf{X}_n$ , where

$$\mathbf{X}_n := \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \quad \text{and} \quad \mathbf{Y}_n := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

assuming that the following model is suitable:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

where  $\epsilon_1, \dots, \epsilon_n \in \mathbb{R}$  and  $\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$  are  $n+1$  conditionally independent random variables given  $\sigma$  with a conditional density for  $\epsilon_i$  given by

$$\epsilon_i | \boldsymbol{\beta}, \sigma \stackrel{\mathcal{D}}{=} \epsilon_i | \sigma \stackrel{\mathcal{D}}{\sim} \frac{1}{\sigma} f\left(\frac{\epsilon_i}{\sigma}\right), \quad i = 1, \dots, n.$$

- (ii) We assume that  $f$  is a strictly positive continuous probability density function on  $\mathbb{R}$  that is symmetric with respect to the origin, and that is such that both tails of  $|z|f(z)$  are monotonic (see Section 5 for the detailed definition of monotonicity), which implies that the tails of  $f(z)$  are also monotonic. Note that these assumptions on  $f$  imply that  $f(z)$  and  $|z|f(z)$  are bounded on the real line, with a limit of 0 in their tails as  $|z| \rightarrow \infty$ . The density  $f$  can have parameters, e.g. a shape parameter; however, their value is assumed to be known.
- (iii) We assume that the joint prior density of  $\boldsymbol{\beta}$  and  $\sigma$ , given by  $\pi(\boldsymbol{\beta}, \sigma)$ , is such that  $\min(\sigma, 1)\pi(\boldsymbol{\beta}, \sigma)$  is bounded (or equivalently that  $\pi(\boldsymbol{\beta}, \sigma)/\max(1, 1/\sigma)$  is bounded). Note that, if we have no prior information, we can set  $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ , the usual non-informative prior for this type of random variables, or simply  $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ .

From this perspective,  $\mathbf{x}_1, \dots, \mathbf{x}_n$  represent observations of the explanatory variables, the dependent variable and the error term are respectively represented by the continuous random variables  $Y_1, \dots, Y_n$  and  $\epsilon_1, \dots, \epsilon_n$ , and the parameters  $\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^T$  and  $\sigma$  represent the regression coefficients and the scale parameter, respectively. Inference on these parameters helps us study the relationship between the explanatory variables and the dependent variable. Note that no assumptions are made on the explanatory variables; they can be continuous, discrete, with any distributions.

Given that the scale parameter of the distribution of the error term is  $\sigma$ , homoscedasticity is an underlying assumption of the model. When the classical framework is considered, i.e. a frequentist setting with the assumption that  $f$  is the standard normal density,  $\sigma$  also represents the standard deviation of each error  $\epsilon_i$ . In this situation, the maximum likelihood estimator of  $\boldsymbol{\beta}$  is the well known least square estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{y}_n$  (provided that  $\mathbf{X}_n^T \mathbf{X}_n$  is invertible).

An important drawback of the classical framework is that outliers have a significant impact on the estimation, due to the normal assumption. In this paper, we study robustness of the estimation of  $\boldsymbol{\beta}$  and  $\sigma$ . The objective is to find sufficient conditions to attain whole robustness, meaning a gradual decrease in the impact of outliers as they approach plus or minus infinity, to ultimately reach a level where their impact is null. The nature of the results presented in Sect. 2.3 is asymptotic, in the sense that some  $y_i$ 's approach  $+\infty$  or  $-\infty$ . In our theoretical analysis, we consider the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n$  as fixed. Studying this theoretical framework is sufficient to attain, in practise, robustness against any type of outliers (i.e. outliers because of their extreme  $\mathbf{x}$  value, extreme  $y$  value, or both). Indeed, an outlier with an extreme  $\mathbf{x}$  value can be viewed as an observation with a fixed  $\mathbf{x}$  value and an extreme  $y$  value, compared with the trend emerging from the bulk of the data.

We now set our mathematical framework for dealing with outliers. Among the  $n$  observations of  $Y_1, \dots, Y_n$ , we assume that  $k > p + 1$  of them, denoted by the vector  $\mathbf{y}_k$ , form a group of nonoutlying observations,  $l$  of them are considered as “lower outliers” with relatively small values of  $y_i$ , and  $u$  of them are considered as “upper outliers” with relatively large values of  $y_i$ , with  $k + l + u = n$ . For  $i = 1, \dots, n$ , we define the binary functions  $k_i, l_i$  and  $u_i$  as follows: if  $y_i$  is a nonoutlying value,  $k_i = 1$ ; if it is a lower outlier,  $l_i = 1$  and if it is an upper outlier,  $u_i = 1$ . These functions take the value of 0 otherwise. Therefore, we have  $k_i + l_i + u_i = 1$  for  $i = 1, \dots, n$ , with  $\sum_{i=1}^n k_i = k$ ,  $\sum_{i=1}^n l_i = l$  and  $\sum_{i=1}^n u_i = u$ . We assume that each outlier converges towards  $-\infty$  or  $+\infty$  at its own specific rate, to the extent that the ratio of two outliers is bounded. More precisely, we assume that  $y_i = a_i + b_i \omega$ , for  $i = 1, \dots, n$ , where  $a_i$  and  $b_i$  are constants such that  $a_i \in \mathbb{R}$  and

$$(i) \ b_i = 0 \text{ if } k_i = 1, \quad (ii) \ b_i < 0 \text{ if } l_i = 1, \quad (iii) \ b_i > 0 \text{ if } u_i = 1,$$

and we let  $\omega \rightarrow \infty$ .

Let the joint posterior density of  $\boldsymbol{\beta}$  and  $\sigma$  be denoted by  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$  and the marginal density of  $(Y_1, \dots, Y_n)$  be denoted by  $m(\mathbf{y}_n)$ , where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = [m(\mathbf{y}_n)]^{-1} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0.$$

Let the joint posterior density of  $\boldsymbol{\beta}$  and  $\sigma$  arising from the nonoutlying observations only be denoted by  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  and the corresponding marginal density be denoted by  $m(\mathbf{y}_k)$ , where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) = [m(\mathbf{y}_k)]^{-1} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[ (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0.$$

Note that if the prior  $\pi(\boldsymbol{\beta}, \sigma)$  is proportional to 1, the likelihood functions, given by the product terms in the posteriors above, can also be expressed as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = m(\mathbf{y}_n) \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) \quad \text{and} \quad \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) = m(\mathbf{y}_k) \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k). \quad (1)$$

**Proposition 1.** *Consider the Bayesian context given in Section 2.1 and assume that the matrix with  $k$  rows given by  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}$  has full rank, where  $k_{i_1} = \dots = k_{i_k} = 1$  ( $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}$  are the observations of the explanatory variables related to nonoutliers). Then, the joint posterior densities  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  and  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$  are proper.*

The proof of Proposition 1 is given in Sect. 5.1. Note that the assumption of Proposition 1 is satisfied when the explanatory variables are continuous.

## 2.2. Log-Regularly Varying Distributions

As mentioned in the introduction, our approach to attain robustness is to replace the traditional normal assumption on the error term by a log-regularly varying distribution assumption. The definition of such a distribution is now presented.

**Definition 1** (Log-regularly varying distribution). *A random variable  $Z$  with a symmetric density  $f(z)$  is said to have a log-regularly varying distribution with index  $\rho \geq 1$  if  $zf(z) \in L_\rho(\infty)$ , meaning that  $zf(z)$  is log-regularly varying at  $\infty$  with index  $\rho \geq 1$  (see Definition 2).*

**Definition 2** (Log-regularly varying function). *We say that a measurable function  $g$  is log-regularly varying at  $\infty$  with index  $\rho \in \mathbb{R}$  if*

$$\begin{aligned} &\forall \epsilon > 0, \forall \tau \geq 1, \text{ there exists a constant } A(\epsilon, \tau) > 0 \text{ such that} \\ &z \geq A(\epsilon, \tau) \text{ and } 1/\tau \leq \nu \leq \tau \Rightarrow |\nu^\rho g(z^\nu)/g(z) - 1| < \epsilon. \end{aligned}$$

*If  $\rho = 0$ ,  $g$  is said to be log-slowly varying at  $\infty$ .*

Log-regularly varying functions is an interesting class of functions with useful properties for robustness. As indicated in Definition 2, they are such that  $g \in L_\rho(\infty)$  if  $g(z^\nu)/g(z)$  converges towards  $\nu^{-\rho}$  uniformly in any set  $\nu \in [1/\tau, \tau]$  (for any  $\tau \geq 1$ ) as  $z \rightarrow \infty$ , where  $\rho \in \mathbb{R}$ . This implies that for any  $\rho \in \mathbb{R}$ , we have  $g \in L_\rho(\infty)$  if and only if there exists a constant  $A > 1$  and a function  $s \in L_0(\infty)$  such that for  $z \geq A$ ,  $g$  can be written as  $g(z) = (\log z)^{-\rho} s(z)$ . It gives you an overview of the tail behaviour of log-regularly varying distributions. Note that an example of such a distribution is presented in Sect. 3.1. For more information on log-regularly varying distributions and log-regularly varying functions, we refer the reader to [Desgagné \(2013\)](#) and [Desgagné \(2015\)](#).

### 2.3. Resolution of conflicts

We now give in Theorem 1 robustness results for models using auxiliary information, i.e. for models with  $p \geq 2$ . Theorem 1 represents the main contribution of this paper. Note that if  $p = 1$ , the linear regression model becomes the location-scale model, and we refer the reader to [Desgagné \(2015\)](#) for suitable assumptions under which robustness is attained in this specific situation.

**Theorem 1.** *Consider the model and the context described in Section 2.1. If we assume that*

- (i)  $zf(z) \in L_\rho(\infty)$  with  $\rho \geq 1$  (i.e. that  $f$  is a log-regularly varying distribution),
- (ii) *there exists  $(y_{i_1}, \dots, y_{i_{l+u+2p-1}}) \subset \mathbf{y}_k$  such that any  $p$  vectors picked among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{l+u+2p-1}}$  are linearly independent,*

*then we obtain the following results:*

(a)

$$\lim_{\omega \rightarrow \infty} \frac{m(\mathbf{y}_n)}{\prod_{i=1}^n [f(y_i)]^{l_i+u_i}} = m(\mathbf{y}_k),$$

(b)

$$\lim_{\omega \rightarrow \infty} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k),$$

*uniformly on  $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$ , for any  $\lambda \geq 0$  and  $\tau \geq 1$ ,*

(c)

$$\lim_{\omega \rightarrow \infty} \int_0^\infty \int_{\mathbb{R}^p} |\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) - \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)| d\boldsymbol{\beta} d\sigma = 0,$$

(d) *As  $\omega \rightarrow \infty$ ,*

$$\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \boldsymbol{\beta}, \sigma \mid \mathbf{y}_k,$$

*and in particular*

$$\beta_i \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \beta_i \mid \mathbf{y}_k, i = 1, \dots, p, \quad \text{and} \quad \sigma \mid \mathbf{y}_n \xrightarrow{\mathcal{D}} \sigma \mid \mathbf{y}_k,$$

(e)

$$\lim_{\omega \rightarrow \infty} [m(\mathbf{y}_k)/m(\mathbf{y}_n)] \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) = \mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k),$$

*uniformly on  $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$ , for any  $\lambda \geq 0$  and  $\tau \geq 1$ .*

The proof of Theorem 1 is given in Sect. 5.2.

Theorem 1 is of great practical use considering the simplicity of its two sufficient conditions. Indeed, condition (i) indicates that modelling must be done using a density  $f$  with sufficiently heavy tails. More precisely,  $f$  must be a log-regularly varying distribution (see Definition 1). Well known heavy-tailed distributions, such as the Student distribution, do not satisfy this criterion. Desgagné (2015) introduced an appealing distribution family that belongs to the family of log-regularly varying distributions, and therefore, satisfies condition (i): the family of log-Pareto-tailed symmetric distributions. In our opinion, the most appealing log-Pareto-tailed symmetric distribution is a distribution called the log-Pareto-tailed standard normal distribution, with parameters  $\alpha > 1$  and  $\phi > 1$ . It exactly matches the standard normal on the interval  $[-\alpha, \alpha]$ , while having tails that behave like  $(1/|z|)(\log |z|)^{-\phi}$  (which is a log-Pareto behaviour). We use this distribution and describe it in our numerical study in Sect. 3.

Condition (ii) indicates that there must be at least  $l + u + 2p - 1$  nonoutliers, say  $y_{i_1}, \dots, y_{i_{l+u+2p-1}}$ , such that if you select any  $p$  vectors among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{l+u+2p-1}}$ , for instance  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$ , they must be linearly independent. Therefore, if there is a relatively large number of different observations of the explanatory variables in a data set (comprised of all the observations), condition (ii) should be satisfied given the usual rarity of outliers (which implies that  $l + u + 2p - 1$  should be relatively close to  $2p - 1$ ). To fix ideas, consider a sample of size  $n = 30$  with one explanatory variable (the simple linear regression model is therefore used, i.e.  $p = 2$ ), comprised of 20 different observations of the explanatory variable. If there are 2 outliers for instance, there are at least 18 different observations of the explanatory variable related to nonoutliers (and  $18 \geq l + u + 2p - 1 = 5$ ); condition (ii) is therefore satisfied. Theorem 1 is as a result particularly well suited for discrete or continuous explanatory variables. Note that when the explanatory variables are continuous, condition (ii) becomes:  $k \geq l + u + 2p - 1$  (because all observations of the explanatory variables are different), i.e. that the difference between the number of nonoutliers and the number of outliers must be greater than or equal to  $2p - 1$ . For instance, with a sample of size  $n = 15$ , robustness is attained provided that there are  $l + u = 6$  outliers (which leaves  $k = 9$  nonoutliers), or less, considering one continuous covariate (therefore using simple linear regression).

Another feature of Theorem 1 that contributes to its practical use is that the results are easy to interpret. In result (a), the asymptotic behaviour of the marginal  $m(\mathbf{y}_n)$  is described. Result (a) is the centrepiece; it is the result that is difficult to prove and it leads to result (b), which indicates that the posterior density, arising from the whole sample, converges towards the posterior density arising from the nonoutliers only, uniformly in any set  $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$ . The impact of outliers then gradually vanishes as they approach plus or minus infinity. Result (b) leads to result (c): the convergence in  $L_1$  of the posterior density, arising from the whole sample, towards the posterior density arising from the nonoutlying observations only. This last result implies the following convergence:  $\mathbb{P}(\boldsymbol{\beta}, \sigma \in E \mid \mathbf{y}_n) \rightarrow \mathbb{P}(\boldsymbol{\beta}, \sigma \in E \mid \mathbf{y}_k)$  as  $\omega \rightarrow \infty$ , uniformly for all sets  $E \subset \mathbb{R}^p \times \mathbb{R}^+$ . This result is slightly stronger than convergence in distribution (result (d)) which requires only pointwise convergence. The convergence of the posterior marginal distributions then follows. Result (d) is the most practical result; it indicates that any estimation of  $\boldsymbol{\beta}$  and  $\sigma$  based on posterior quantiles (e.g. using posterior medians and Bayesian credible intervals) is robust to outliers. Result (e), which follows from result (b), indicates that, for a given sample, the likelihood (up to a multiplicative constant that does not depend on  $\boldsymbol{\beta}$  and  $\sigma$ ) converges to the likelihood



arising from the nonoutliers only, uniformly in any set  $(\boldsymbol{\beta}, \sigma) \in E$ , where  $E = [-\lambda, \lambda]^p \times [1/\tau, \tau]$ . Consequently, the maximum of  $\mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$  thus converges to the maximum of  $\mathcal{L}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  on the set  $E$ , and therefore, the maximum likelihood estimate also converges, as  $\omega \rightarrow \infty$ .

### 3. Numerical Study

This section is dedicated to practical considerations related to Theorem 1. More precisely, in Sect. 3.1, we illustrate the practical implications of the theoretical results of Theorem 1 via an analysis of the relationship between two stock market indexes. Through this analysis, we also present our simple Bayesian approach to robustly identify statistically significant linear dependencies between variables. In Sect. 3.2, a simulation study is conducted to evaluate the accuracy of the estimates arising from a robust multiple linear regression model, based on our approach. In all the analyses, we compare the results arising from our robust model with those of the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models.

#### 3.1. Analysis of the Relationship Between S&P 500 and S&P/TSX

It is well known that the returns of the S&P 500 and S&P/TSX are positively correlated. We first illustrate through analyses of returns of these two stock market indexes that when we artificially move an observation, its impact on the estimation grows until it reaches a certain threshold. Beyond this threshold, the impact vanishes as the observation converges towards plus or minus infinity. For the analyses, we consider the monthly returns from beginning of October 2014 to beginning of September 2016, and that the explanatory and dependent variables are the returns of the S&P/TSX and S&P 500, respectively. The data are presented in Table 1.

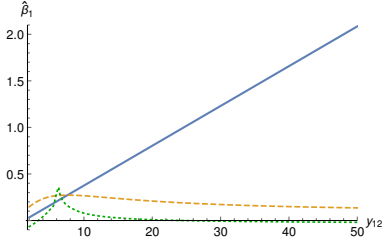
$y_i$	-0.12	-0.12	3.56	0.09	1.53	0.27	6.60	-0.41	-5.07	-1.75	0.05	$y_{12}$
$x_i$	0.88	0.10	3.68	-0.01	0.82	3.39	4.93	0.30	-1.44	-3.41	-0.44	1.67
$y_i$	-2.64	-6.26	1.97	-2.10	1.05	0.85	-1.74	5.49	-3.10	-0.42	2.45	
$x_i$	-3.98	-4.21	-0.58	-3.07	-1.38	2.16	-2.18	3.82	0.28	-0.76	0.90	

Table 1: Returns for month  $i$  of S&P 500 ( $y_i$ ) and S&P/TSX ( $x_i$ ) in percentages, for  $i = 1, \dots, 23$

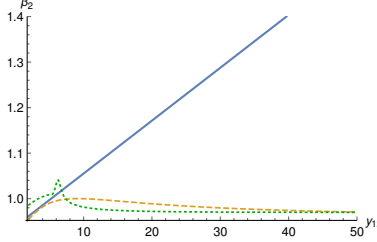
In order to illustrate the threshold feature, an observation is randomly chosen (in this analysis, it is the 12th observation), and  $y_{12}$  is gradually moved from the value 2 (a nonoutlier) to 50 (a large outlier), while  $x_{12} = 1.67$  remains fixed. Note that the true value of  $y_{12}$  is 8.30. The parameters  $\boldsymbol{\beta} := (\beta_1, \beta_2)$  and  $\sigma$  are estimated for each data set related to a different value of  $y_{12}$  using maximum *a posteriori* probability (MAP) estimation with a prior proportional to 1 (in this situation, MAP estimation corresponds to maximum likelihood estimation). This process is performed under three models, each corresponding to a different assumption on  $f$ : a standard normal density (in this case,  $\hat{\boldsymbol{\beta}}$  is the classical least square estimator), a Student density (the partially robust model) or a log-Pareto-tailed standard normal density (our robust model, see (2) for the distribution). The results are presented in Figure 1.



Estimation of  $\beta_1$  under the three models when  $y_{12}$  increases from 2 to 50



Estimation of  $\beta_2$  under the three models when  $y_{12}$  increases from 2 to 50



Estimation of  $\sigma$  under the three models when  $y_{12}$  increases from 2 to 50

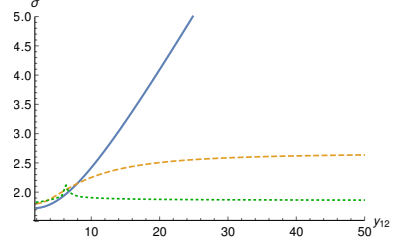


Figure 1: Estimation of  $\beta$  and  $\sigma$  when  $y_{12}$  increases from 2 to 50 under three different assumptions on  $f$ : standard normal density (blue solid line), Student density (orange dashed line) and log-Pareto-tailed standard normal density (green dotted line)

The inference is clearly not robust when it is assumed that the error has a normal distribution, because the estimates of  $\beta_1$ ,  $\beta_2$  and  $\sigma$  increase with  $y_{12}$ . For the partially robust model, the degrees of freedom of the heavy-tailed Student distribution have been arbitrarily set to 10 and a known scale parameter of 0.88 has been added to this distribution in order to have the same 2.5th and 97.5th percentiles as the standard normal. The estimation of  $\beta$  is robust as the impact of the outlier slowly decreases after a certain threshold. However, the estimation of  $\sigma$  is only partially robust, i.e. the impact of the outlier is limited, but does not decrease when the outlying value increases. For our robust model, we have assumed that the error term has a log-Pareto-tailed standard normal distribution. We have arbitrarily set  $\alpha = 1.96$ , which implies that, according to the procedure described in Section 4 of [Desgagné \(2015\)](#),  $\phi = 4.08$  (this procedure ensures that  $f$  is continuous and a probability density function). Therefore, all three distributions studied in this section have 95% of their mass in the interval  $[-1.96, 1.96]$ . The density of the log-Pareto-tailed standard normal distribution is given by

$$f(x) = \begin{cases} (2\pi)^{-1/2} \exp(-x^2/2) & \text{if } |x| \leq \alpha \text{ (the standard normal part),} \\ (2\pi)^{-1/2} \exp(-\alpha^2/2)(\alpha/|x|)(\log \alpha / \log |x|)^\phi & \text{if } |x| > \alpha \text{ (the log-Pareto tails),} \end{cases} \quad (2)$$

and depicted in Figure 2.

### Comparison between the standard normal, Student and log-Pareto-tailed standard normal

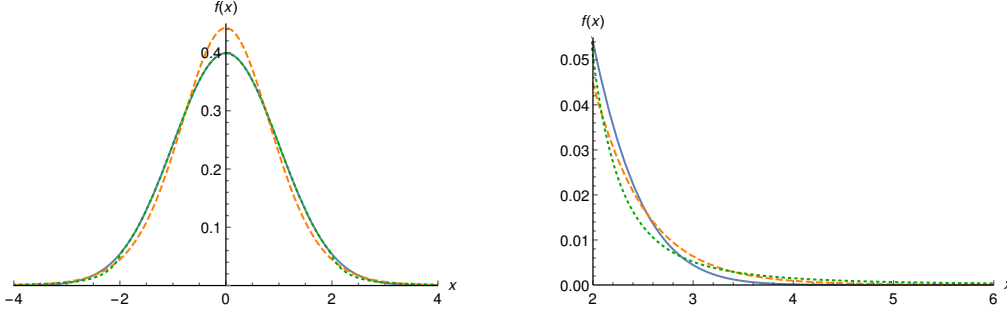


Figure 2: Densities of the standard normal (blue solid line), Student with 10 degrees of freedom and a known scale parameter of 0.88 (orange dashed line), and log-Pareto-tailed standard normal with  $\alpha = 1.96$  and  $\phi = 4.08$  (green dotted line)

For our robust model, it can be seen that  $y_{12}$  has an increasing impact on the estimation until this observation reaches a threshold. In this analysis, the threshold is around  $y_{12} = 6$ , and based on the data set with  $y_{12} = 6$ ,  $\hat{\beta}_1 = 0.37$ ,  $\hat{\beta}_2 = 1.05$  and  $\hat{\sigma} = 2.13$ . Beyond this threshold, the impact of the outlier gradually vanishes as it converges towards infinity. The point estimates converge towards  $-0.06$ ,  $0.98$  and  $1.84$  for  $\beta_1$ ,  $\beta_2$  and  $\sigma$ , respectively, which are the point estimates when  $(x_{12}, y_{12})$  is excluded from the sample. Whole robustness is therefore attained for both  $\beta$  and  $\sigma$ . Note that an increase in the value of the parameter  $\alpha$  would result in an increase in the value of the threshold. Setting  $\alpha = 1.96$  seems to be suitable for practical use.

We now present a more typical analysis of the relationship between S&P 500 and S&P/TSX, based on the original data set, i.e. with  $y_{12} = 8.30$ . Afterwards, we present an analysis of a more atypical data set of returns to illustrate the relevance of our approach in analyses of the relationship between two variables. The parameter estimates are presented in Table 2. They are relatively similar for the three models. They indicate that the parameter  $\beta_1$  might not be of great importance for the models (since the posteriors have significant mass around 0), and that it might be a statistically significant linear dependence between S&P/TSX and S&P 500 (since the posteriors do not have significant mass around 0). We confirm the linear dependency through a Bayesian test for  $H_0 : \beta_2 = 0$  versus  $H_1 : \beta_2 \neq 0$  using the following Bayes factor:  $m(\mathbf{y}_n)/m(\mathbf{y}_n | H_0)$  (i.e. the marginal density of  $\mathbf{y}_n$  divided by the marginal density of  $\mathbf{y}_n$  under  $H_0$ , which is the marginal density of  $\mathbf{y}_n$  under the simple location-scale model). We consider the table provided by Kass and Raftery (1995) as guidelines and present it in Table 3. The Bayes factors for all the three models are larger than 150. Therefore, there is clearly enough statistical evidence to confirm the linear dependency between S&P/TSX and S&P 500.

Assumptions on $f$	Estimates for		
	$\beta_1$	$\beta_2$	$\sigma$
Standard normal	0.30 (−0.72, 1.32)	1.04 (0.62, 1.45)	2.39 (1.72, 3.26)
Student (10 d.f.)	0.27 (−0.70, 1.23)	1.00 (0.61, 1.40)	2.38 (1.62, 3.38)
Log-Pareto-tailed normal	0.16 (−0.80, 1.21)	1.01 (0.62, 1.41)	2.24 (1.53, 3.23)

Table 2: Posterior medians and 95% highest posterior density (HPD) intervals under the three models, based on the analysis of the data set presented in Table 1 with  $y_{12} = 8.30$

Bayes factor	Evidence against $H_0$
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Table 3: Table provided by [Kass and Raftery \(1995\)](#) as guidelines for Bayes factors

An analysis of the daily returns of January 2011 (the data are presented in Table 4) leads to different conclusions when the nonrobust model is used, as shown in Figure 3 and Table 5. Indeed, the 95% HPD interval of  $\beta_2$  now includes 0, which suggests that it might not be a statistically significant linear dependence between S&P/TSX and S&P 500. In addition, the Bayes factor for testing  $H_0 : \beta_2 = 0$  versus  $H_1 : \beta_2 \neq 0$  is 3.41, which corresponds to, rounded up to the nearest integer, the upper bound of the interval representing a “not worth more than a bare mention” type of evidence against  $H_0$ . Considering the analysis based on the nonrobust model, one could conclude that there is not enough statistical evidence to confirm the linear dependency between S&P/TSX and S&P 500. Note that the Bayes factors for the partially robust and robust models are 7.69 and 17.08, respectively.

We observe the presence of one clear outlier:  $(x_{18}, y_{18}) = (0.20, -1.79)$  (because of its extremely low  $y$  value, especially compared with the trend emerging from the bulk of the data). In order to evaluate the impact of this outlier and draw conclusions based on the bulk of the data, we redo the analysis while excluding observation 18. The results are showed in Figure 4 and Table 6. The results are now relatively similar for the three models and, as in the analysis of the monthly returns, they indicate that it might be a statistically significant linear dependence between S&P/TSX and S&P 500. This is confirmed by the Bayes factors, which are 47.35, 39.45 and 44.16 for the nonrobust, partially robust and robust models, respectively.

The inference arising from our robust model, based on the original data set of the daily returns of January 2011, is the one that best reflects the behaviour of the bulk of the data, compared to the inferences arising from the nonrobust and partially robust models. Our robust model therefore succeeds in limiting the influence of outliers in order to obtain conclusions consistent with the majority of the observations. We finally note that our robust model also leads to a simple Bayesian approach to robustly identify a statistically significant linear dependence between two variables: assume that the error term has a log-Pareto-tailed standard normal density with parameters  $\alpha = 1.96$  and  $\phi = 4.08$ , and if the 95% HPD interval of  $\beta_2$  does not include 0 and the Bayes factor for

testing  $H_0 : \beta_2 = 0$  versus  $H_1 : \beta_2 \neq 0$  is relatively far from 3, then conclude that there is enough statistical evidence to confirm the linear dependency. This procedure is also valid to evaluate the strength of the relationship between the dependent variable and explanatory variables in a context of multiple linear regression.

$y_i$	-0.13	0.50	-0.21	-0.18	-0.14	0.37	0.90	-0.17	0.74	0.14
$x_i$	-0.30	-0.05	-0.63	-0.30	-0.20	1.18	0.44	-0.44	0.47	0.71
$y_i$	-1.01	-0.13	0.24	0.58	0.03	0.42	0.22	-1.79	0.77	
$x_i$	-0.89	-0.80	-0.55	0.67	-0.66	1.56	-0.41	0.20	0.85	

Table 4: Returns for day  $i$  of July 2011 of S&P 500 ( $y_i$ ) and S&P/TSX ( $x_i$ ) in percentages, for  $i = 1, \dots, 19$

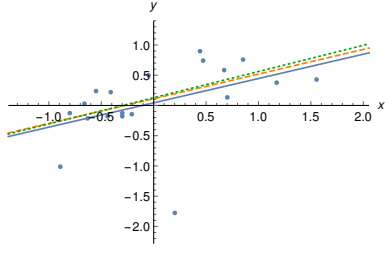
Assumptions on $f$	Estimates for		
	$\beta_1$	$\beta_2$	$\sigma$
Standard normal	0.04 (-0.25, 0.34)	0.40 (-0.02, 0.83)	0.62 (0.43, 0.88)
Student (10 d.f.)	0.11 (-0.14, 0.35)	0.41 (0.07, 0.76)	0.53 (0.33, 0.81)
Log-Pareto-tailed normal	0.13 (-0.09, 0.34)	0.43 (0.13, 0.72)	0.42 (0.26, 0.65)

Table 5: Posterior medians and 95% highest posterior density (HPD) intervals under the three models based on the analysis of the data set presented in Table 4

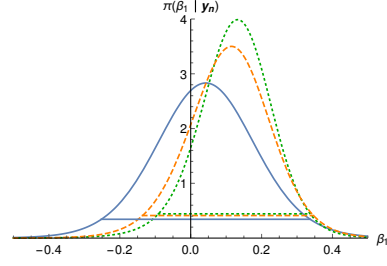
Assumptions on $f$	Estimates for		
	$\beta_1$	$\beta_2$	$\sigma$
Standard normal	0.15 (-0.03, 0.33)	0.44 (0.18, 0.69)	0.37 (0.25, 0.53)
Student (10 d.f.)	0.16 (-0.02, 0.34)	0.41 (0.16, 0.68)	0.39 (0.25, 0.58)
Log-Pareto-tailed normal	0.15 (-0.04, 0.33)	0.43 (0.17, 0.69)	0.37 (0.24, 0.54)

Table 6: Posterior medians and 95% highest posterior density (HPD) intervals under the three models based on the analysis of the data set presented in Table 4, but excluding observation 18

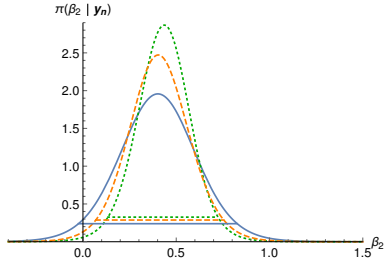
(a) Daily returns for July 2011 of S&P 500 and S&P/TSX



(b) Posterior density of  $\beta_1$



(c) Posterior density of  $\beta_2$



(d) Posterior density of  $\sigma$

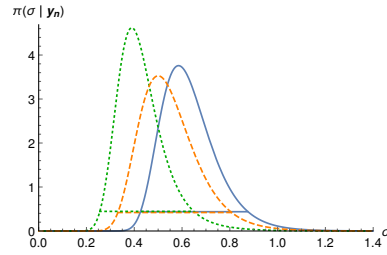
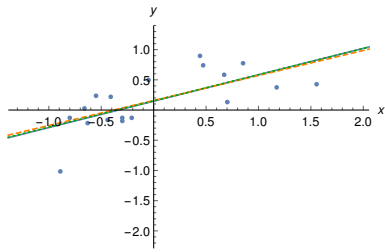
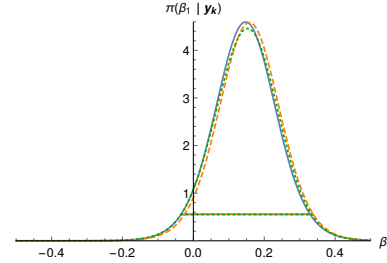


Figure 3: (a) Daily returns for July 2011 of S&P 500 and S&P/TSX, (b)-(d) Posterior densities of  $\beta_1$ ,  $\beta_2$  and  $\sigma$  arising from the original data set with 95% HPD intervals (horizontal lines); for each graph, the blue solid, orange dashed and green dotted lines are respectively related to the nonrobust, partially robust and robust models

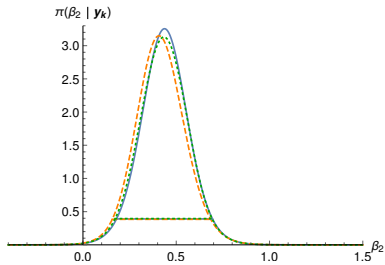
(a) Daily returns for July 2011 of S&P 500 and S&P/TSX excluding observation 18



(b) Posterior density of  $\beta_1$  excluding observation 18



(c) Posterior density of  $\beta_2$  excluding observation 18



(d) Posterior density of  $\sigma$  excluding observation 18

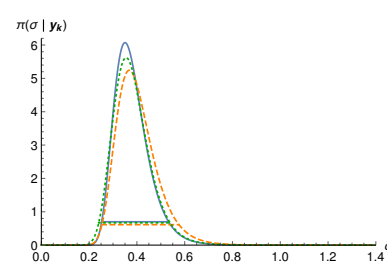


Figure 4: (a) Daily returns for July 2011 of S&P 500 and S&P/TSX excluding observation 18, (b)-(d) Posterior densities of  $\beta_1$ ,  $\beta_2$  and  $\sigma$  arising from the data set excluding observation 18 with 95% HPD intervals (horizontal lines); for each graph, the blue solid, orange dashed and green dotted lines are respectively related to the nonrobust, partially robust and robust models

### 3.2. Simulation Study

In this section, we evaluate through a simulation study the accuracy of the estimates arising from a robust multiple linear regression model with two explanatory variables, based on our approach. More precisely, the model is:  $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$  with  $\boldsymbol{\beta} := (\beta_1, \beta_2, \beta_3)$  and  $\epsilon_i \mid \sigma \stackrel{\mathcal{D}}{\sim} (1/\sigma)f(\epsilon_i/\sigma)$ ,  $i = 1, \dots, n$ , where  $f$  is assumed to be a log-Pareto-tailed standard normal density with  $\alpha = 1.96$  and  $\phi = 4.08$  in our robust model, and it is compared with the same linear regression model, but where  $f$  is assumed to be a standard normal density in the nonrobust model, and a Student density with 10 degrees of freedom and a known scale parameter of 0.88 in the partially robust model. Note that the parameters of the distributions are the same as in Sect. 3.1.

For the simulation study, we set  $n = 30$ ,  $x_{1,2}, x_{2,2}, \dots, x_{30,2} = 1, 2, \dots, 30$  (which are the observations of the first explanatory variable),  $x_{1,3}, x_{2,3}, \dots, x_{30,3} = 0^2, 1^2, \dots, 29^2$  (which are the observations of the second explanatory variable), and  $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ . We simulate 50,000 data sets using values for  $\boldsymbol{\beta}$  and  $\sigma$  arbitrarily set to  $(10, 1, -0.1)$  and 2, respectively, and we carry out this process for each of the three scenarios that we now describe. In the first one,  $f$  is a standard normal distribution; therefore, the probability to observe outliers is negligible. In the second scenario,  $f$  is a mixture of two normals where the first component is a standard normal distribution and the second has a mean of 0 and a variance of  $10^2$ , with weights of 0.9 and 0.1, respectively. This last component can contaminate the data sets by generating extreme values. In the third and last scenario,  $f$  is also a mixture of two normals, but the contamination is due to the second component's location. More precisely, the first component is again a standard normal, but the second has a mean of 10 and a variance of 1, with weights of 0.95 and 0.05, respectively.

For each simulated data set, we estimate  $\boldsymbol{\beta}$  and  $\sigma$  using MAP estimation for the three models. Then, within each simulation scenario, we evaluate the performance of each model using sample mean square errors (MSE), based on the true values  $\beta_1 = 10, \beta_2 = 1, \beta_3 = -0.1$  and  $\sigma = 2$ , where the performance of the estimation  $\boldsymbol{\beta}$  is evaluated through the sum of the MSE of the estimators of  $\beta_1, \beta_2$  and  $\beta_3$ . The results are presented in Tables 7 and 8.

We first notice that our robust model performs best even when there is no outliers (the 100%  $\mathcal{N}(0, 1)$  scenario). We believe that this is explained by the size of the samples. Some of them being nonrepresentative, the robust model is still able to detect the appropriate trend. For the two other scenarios, it is worse for the nonrobust and partially robust models because of the frequent presence of outliers. For the robust model, we observe that outliers have minimal impact on the estimation of both  $\boldsymbol{\beta}$  and  $\sigma$ , which confirms that the proposed approach provides whole robustness with respect to outliers.

Assumptions on $f$	Scenarios		
	100% $\mathcal{N}(0, 1)$	90% $\mathcal{N}(0, 1)$ + 10% $\mathcal{N}(0, 10^2)$	95% $\mathcal{N}(0, 1)$ + 5% $\mathcal{N}(10, 1)$
Standard normal	1.40	15.15	8.99
Student (10 d.f.)	1.11	3.37	2.76
Log-Pareto-tailed normal	0.18	0.39	0.30

Table 7: Sum of the MSE of the estimators of  $\beta_1, \beta_2$  and  $\beta_3$  under the three scenarios and the three assumptions of  $f$

Assumptions on $f$	Scenarios		
	$100\%N(0, 1)$	$90\%N(0, 1) + 10\%N(0, 10^2)$	$95\%N(0, 1) + 5\%N(10, 1)$
Standard normal	0.08	20.61	7.68
Student (10 d.f.)	0.07	5.16	2.56
Log-Pareto-tailed normal	0.05	0.21	0.08

Table 8: MSE of the estimators of  $\sigma$  under the three scenarios and the three assumptions of  $f$

#### 4. Conclusion

In this paper, we have provided a simple approach to attain robustness against outliers in Bayesian linear regression: replace the traditional normal assumption on the error term by a super heavy-tailed distribution assumption. We have proved that whole robustness is attained provided that there exists  $(y_{i_1}, \dots, y_{i_{l+u+2p-1}}) \subset \mathbf{y}_k$  such that any  $p$  vectors picked among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{l+u+2p-1}}$  are linearly independent, as stated in Theorem 1. In Sect. 3.2, we have evaluated the performance of our approach through a simulation study. We have also explained in Sect. 3.1 that our approach leads to a simple procedure to robustly identify a statistically significant linear dependence between two variables.

The robust model that we have used in all the numerical analyses was based on the assumption that the error term had the log-Pareto-tailed standard normal density given in (2). This robust model has been compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models. The conclusion is: our robust model performs as well as the nonrobust and the partially robust models in absence of outliers, in addition to being completely robust. Therefore, our recommendation is the following: assume that the error has the density given in (2) and obtain adequate results, regardless of whether there are outliers, by computing estimates as usual from the posterior distribution.

#### 5. Proofs

The proof of Proposition 1 is given in Section 5.1 and the proof of Theorem 1 is given in Section 5.2. Beforehand, we define the constant  $B > 0$  as follows:

$$B = \max \left\{ \sup_{z \in \mathbb{R}} f(z), \sup_{z \in \mathbb{R}} |z|f(z), \sup_{\beta \in \mathbb{R}^p, \sigma > 0} \min(\sigma, 1)\pi(\beta, \sigma) \right\}.$$

The monotonicity of the tails of  $f(z)$  and  $|z|f(z)$  implies that there exists a constant  $M > 0$  such that

$$|y| \geq |z| \geq M \Rightarrow f(y) \leq f(z) \text{ and } |y|f(y) \leq |z|f(z). \quad (3)$$

##### 5.1. Proof of Proposition 1

To prove that  $\pi(\beta, \sigma \mid \mathbf{y}_n)$  is proper (the proof for  $\pi(\beta, \sigma \mid \mathbf{y}_k)$  is omitted because it is similar), it suffices to show that the marginal  $m(\mathbf{y}_n)$  is finite. Note that we assume that the matrix with  $k$  rows given by  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}$  has full rank, where  $k_{i_1} = \dots = k_{i_k} = 1$ . This implies that  $\mathbf{X}_n$  has full



rank, which in turns implies that the rank of  $\mathbf{X}_n$  is  $p$  (we assume that  $n \geq k > p + 1$ ), and that there exists  $(i_1, \dots, i_p) \subset (1, \dots, n)$  such that

$$\det \begin{pmatrix} \mathbf{x}_{i_1} \\ \vdots \\ \mathbf{x}_{i_p} \end{pmatrix} \neq 0.$$

We first show that the function is integrable on the area where the ratio  $1/\sigma$  is bounded above. More precisely, we consider  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\delta M^{-1} \leq \sigma < \infty$ , where  $\delta$  is a positive constant. Then, we show that the function is integrable on the area where the ratio  $1/\sigma$  approaches infinity. We assume without loss of generality that the matrix with  $p$  rows given by  $\mathbf{x}_1, \dots, \mathbf{x}_p$  is invertible, and therefore, we have

$$\begin{aligned} & \int_{\delta M^{-1}}^{\infty} \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\boldsymbol{\beta} d\sigma \\ & \stackrel{a}{\leq} B^{n-p+1} \int_{\delta M^{-1}}^{\infty} \max\left(1, \frac{1}{\sigma}\right) \frac{1}{\sigma^{n-p}} \int_{\mathbb{R}^p} \prod_{i=1}^p \frac{1}{\sigma} f\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) d\boldsymbol{\beta} d\sigma \\ & \stackrel{b}{\leq} \max\left(1, \frac{M}{\delta}\right) B^{n-p+1} \left| \det \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_p \end{pmatrix} \right|^{-1} \int_{\delta M^{-1}}^{\infty} \frac{1}{\sigma^{n-p}} \int_{-\infty}^{\infty} f(u_1) du_1 \times \dots \times \int_{-\infty}^{\infty} f(u_p) du_p d\sigma \\ & \propto \int_{\delta M^{-1}}^{\infty} \frac{1}{\sigma^{n-p}} d\sigma \stackrel{c}{=} (M/\delta)^{n-p-1} / (n-p-1) < \infty. \end{aligned}$$

In step *a*, we bound  $\pi(\boldsymbol{\beta}, \sigma) / \max(1, 1/\sigma)$  and each of  $n-p$  densities  $f$  by  $B$ . In step *b*, we use the change of variables  $u_i = (y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma$  for  $i = 1, \dots, p$ . In step *c*, we use  $n > p + 1$ . Note that if instead, in step *a*, we bound  $\sigma\pi(\boldsymbol{\beta}, \sigma)$  by  $B$ , one can verify that the condition  $n \geq p + 1$  is sufficient to bound above the integral.

We now show that the integral is finite on  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $0 < \sigma < \delta M^{-1}$ . In this area, the ratio  $(1/\sigma)$  approaches infinity. We have to carefully analyse the subareas where  $y_i - \mathbf{x}_i^T \boldsymbol{\beta}$  is close to 0 in order to deal with the  $0/0$  form of the ratios  $(y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma$ . In order to achieve this, we split the domain of  $\boldsymbol{\beta}$  into  $p + 1$  mutually exclusive areas as follows:

$$\begin{aligned} \mathbb{R}^p = & \left[ \cap_{i_1=1}^n \mathcal{R}_{i_1}^c \right] \cup \left[ \cup_{i_1=1}^n \left( \mathcal{R}_{i_1} \cap \left( \cap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c \right) \right) \right] \cup \left[ \cup_{i_1, i_2=1(i_1 \neq i_2)}^n \left( \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \left( \cap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c \right) \right) \right] \\ & \cup \dots \cup \left[ \cup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n \left( \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \right) \right], \end{aligned}$$

where  $\mathcal{R}_j := \{\boldsymbol{\beta} : |y_j - \mathbf{x}_j^T \boldsymbol{\beta}| < \delta\}$ ,  $j \in \{1, \dots, n\}$ . The set  $\mathcal{R}_j$  represents hyperplanes passing near the point  $(\mathbf{x}_j, y_j)$ . The set  $\cap_{i_1=1}^n \mathcal{R}_{i_1}^c$  is therefore comprised of hyperplanes passing not near any point. The set  $\cup_{i_1=1}^n (\mathcal{R}_{i_1} \cap (\cap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c))$  represents hyperplanes passing near one (and only one) point. The set  $\cup_{i_1, i_2=1(i_1 \neq i_2)}^n (\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap (\cap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c))$  represents hyperplanes passing near two (and only two) points, and so on.

We choose  $\delta$  small enough to ensure that  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} = \emptyset$  if the vectors  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$  are linearly dependent. To help visualise this, consider that  $\mathbf{x}_{i_p}$  is a linear combination of  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$  for instance, i.e.  $\mathbf{x}_{i_p} = \sum_{j=1}^{p-1} a_j \mathbf{x}_{i_j}$ ,  $a_1, \dots, a_{p-1} \in \mathbb{R}$ , and therefore, if  $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_{p-1}}$  we have  $|y_{i_p} - \mathbf{x}_{i_p}^T \boldsymbol{\beta}| = |y_{i_p} - (\sum_{s=1}^{p-1} a_s \mathbf{x}_{i_s})^T \boldsymbol{\beta}| \approx |y_{i_p} - \sum_{s=1}^{p-1} a_s y_{i_s}| \geq \delta$  (because  $Y_1, \dots, Y_n$  are continuous random variables and the probability of the event  $Y_{i_p} - \sum_{s=1}^{p-1} a_s Y_{i_s} = 0$  is 0). Note that we choose  $\delta$  small enough to ensure that  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} = \emptyset$  if the vectors  $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}$  are linearly dependent,  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \mathcal{R}_{i_3} = \emptyset$  if the vectors  $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}$  are linearly dependent, and so on. Also,  $p+1$  vectors  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p+1}}$  are necessarily linearly dependent (because of the dimension of the space). Therefore,  $\delta$  can be chosen small enough to ensure that  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_{p+1}} = \emptyset$  when  $i_1, \dots, i_{p+1}$  are all different, regardless of the  $\mathbf{x}$  values of the related observations. Therefore, we now know that  $\bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n (\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p})$  (when nonempty) is comprised of hyperplanes passing near  $p$  (and only  $p$ ) points, because  $|y_{i_{p+1}} - \mathbf{x}_{i_{p+1}}^T \boldsymbol{\beta}| \geq \delta$  if  $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}$  for all  $i_{p+1} \notin \{i_1, \dots, i_p\}$ . Note that the set  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}$  is not empty whenever  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$  are linearly independent, and because  $\mathbf{X}_n$  has full rank, we know that there exists at least one nonempty set like this. All this implies that

$$\begin{aligned} & \left[ \bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n (\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}) \right] \\ &= \left[ \bigcup_{i_1, i_2, \dots, i_p=1(i_j \neq i_s \forall i_j, i_s \text{ s.t. } j \neq s)}^n (\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p} \cap (\bigcap_{i_{p+1}=1(i_{p+1} \neq i_1, i_2, \dots, i_p)}^n \mathcal{R}_{i_{p+1}}^c)) \right]. \end{aligned}$$

We now consider  $0 < \sigma < \delta M^{-1}$  and  $\boldsymbol{\beta}$  in one of the subareas given above, i.e.  $\bigcap_{i_1=1}^n \mathcal{R}_{i_1}^c$ ,  $\mathcal{R}_{i_1} \cap (\bigcap_{i_2=1(i_2 \neq i_1)}^n \mathcal{R}_{i_2}^c)$ ,  $\mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap (\bigcap_{i_3=1(i_3 \neq i_1, i_2)}^n \mathcal{R}_{i_3}^c)$ , etc. As explained above, the difficulty lies in dealing with the points  $(\mathbf{x}_i, y_i)$  that are such that  $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| < \delta$ . The strategy is to use them to integrate over  $\boldsymbol{\beta}$ . Therefore, if  $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_p}$ , we use the points  $(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \dots, (\mathbf{x}_{i_p}, y_{i_p})$ . If  $\boldsymbol{\beta} \in \mathcal{R}_{i_1} \cap \mathcal{R}_{i_2} \cap \dots \cap \mathcal{R}_{i_{p-1}} \cap (\bigcap_{i_p=1(i_p \neq i_1, \dots, i_{p-1})}^n \mathcal{R}_{i_p}^c)$ , we use the points  $(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \dots, (\mathbf{x}_{i_{p-1}}, y_{i_{p-1}})$ , and any other point  $(\mathbf{x}_{i_p}, y_{i_p})$  such that the matrix with  $p$  rows given by  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$  is invertible (if  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$  are linearly dependent, the set is empty), and so on. We have

$$\begin{aligned} & \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\ & \stackrel{a}{\leq} (B/\sigma) \max(1, \delta(M)^{-1}) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\ & \propto (1/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \prod_{i \notin \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\ & \stackrel{b}{\leq} (1/\sigma) [(1/\sigma) f(\delta/\sigma)]^{n-p} \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\ & \stackrel{c}{\leq} [B/\delta]^{n-p-1} (1/\sigma^2) f(\delta/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma). \end{aligned}$$

In step *a*, we use  $\pi(\boldsymbol{\beta}, \sigma) \leq \max(\sigma^{-1}, 1)B = \sigma^{-1}B \max(1, \sigma) \leq \sigma^{-1}B \max(1, \delta M^{-1})$ . In step *b*, for all  $i \notin \{i_1, \dots, i_p\}$  we use  $f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \leq f(\delta/\sigma)$  by the monotonicity of the tails of  $f$  because  $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq \delta/\sigma \geq \delta\delta^{-1}M = M$ . In step *c*, we bound  $n - p - 1$  terms  $(1/\sigma)f(\delta/\sigma)$  by  $B/\delta$ .

Finally, we bound the integral of  $(1/\sigma^2)f(\delta/\sigma) \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$  by

$$\begin{aligned} & \int_0^\infty (1/\sigma^2)f(\delta/\sigma) \int_{\mathbb{R}^p} \prod_{i \in \{i_1, \dots, i_p\}} (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\boldsymbol{\beta} d\sigma \\ & \stackrel{a}{=} \left| \det \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_p \end{pmatrix} \right|^{-1} \int_0^\infty (1/\sigma^2)f(\delta/\sigma) d\sigma \stackrel{b}{=} \left| \det \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_p \end{pmatrix} \right|^{-1} \int_0^\infty f(\sigma') d\sigma' < \infty. \end{aligned}$$

In step *a*, we use the same change of variables as above:  $u_j = (y_{i_j} - \mathbf{x}_{i_j}^T \boldsymbol{\beta})/\sigma$  for  $j = 1, \dots, p$ . In step *b*, we use the change of variable  $\sigma' = \delta/\sigma$ .

## 5.2. Proof of Theorem 1

Consider the model and the context described in Section 2.1. We assume that there exists  $(y_{i_1}, \dots, y_{i_{l+u+2p-1}}) \subset \mathbf{y}_k$  such that any  $p$  vectors picked among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{l+u+2p-1}}$  are linearly independent. In addition, we assume that  $l + u \geq 1$ , i.e. that there is at least one outlier, otherwise the proof would be trivial. The proofs of results (a) to (e) are first given, and two propositions and two lemmas that are used in these proofs follow.

*Proof of Result (a).* We first observe that

$$\begin{aligned} \frac{m(\mathbf{y}_n)}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{l_i+u_i}} &= \frac{m(\mathbf{y}_n)}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{l_i+u_i}} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_n) d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[ (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i+l_i+u_i}}{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{l_i+u_i}} d\sigma d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} d\sigma d\boldsymbol{\beta}. \end{aligned}$$

We show that the last integral converges towards 1 as  $\omega \rightarrow \infty$  to prove result (a). If we use Lebesgue's dominated convergence theorem to interchange the limit  $\omega \rightarrow \infty$  and the integral, we have

$$\begin{aligned}
& \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i + u_i} d\sigma d\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \int_0^\infty \lim_{\omega \rightarrow \infty} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i + u_i} d\sigma d\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta} = 1,
\end{aligned}$$

using Proposition 3 in the second equality, since  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are fixed, and then Proposition 1. Note that pointwise convergence is sufficient, for any value of  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\sigma > 0$ , once the limit is inside the integral. However, in order to use Lebesgue's dominated convergence theorem, we need to prove that the integrand is bounded by an integrable function of  $\boldsymbol{\beta}$  and  $\sigma$  that does not depend on  $\omega$ , for any value of  $\omega \geq y$ , where  $y$  is a constant. The constant  $y$  can be chosen as large as we want, and minimum values for  $y$  will be given throughout the proof. In order to bound the integrand, we divide the domain of integration into two areas:  $1 \leq \sigma < \infty$  and  $0 < \sigma < 1$ . Again, we want to separately analyse the area where the ratio  $1/\sigma$  approaches infinity.

We assumed that  $y_i$  can be written as  $y_i = a_i + b_i \omega$ , where  $\omega \rightarrow \infty$ , and  $a_i$  and  $b_i$  are constants such that  $a_i \in \mathbb{R}$  and  $b_i \neq 0$  if  $y_i$  is an outlier. Therefore, the ranking of the elements in the set  $\{|y_i| : l_i + u_i = 1\}$  is primarily determined by the values  $|b_1|, \dots, |b_n|$ , and we can choose the constant  $y$  larger than a certain threshold to ensure that this ranking remains unchanged for all  $\omega \geq y$ . Without loss of generality, we assume for convenience that

$$\omega = \min_{\{i: l_i + u_i = 1\}} |y_i| \quad \text{and consequently} \quad \min_{\{i: l_i + u_i = 1\}} |b_i| = 1.$$

We now bound above the integrand on the first area.

**Area 1:** Consider  $1 \leq \sigma < \infty$  and assume without loss of generality that  $y_1, \dots, y_p$  are  $p$  nonoutliers (therefore  $k_1 = \dots = k_p = 1$ ) such that  $\mathbf{x}_1, \dots, \mathbf{x}_p$  are linearly independent. We have

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) & \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \propto \frac{\pi(\boldsymbol{\beta}, \sigma)}{\sigma^n} \prod_{i=1}^n \frac{f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{[f(y_i)]^{l_i+u_i}} \\
& \stackrel{a}{\leq} \frac{B}{\sigma^n} \prod_{i=1}^n \frac{D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{[f(y_i)]^{l_i+u_i}} \\
& \stackrel{b}{\leq} \frac{1}{[f(\omega)]^{l+u}} \frac{B}{\sigma^n} \prod_{i=1}^n D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) [|b_i|D(|a_i|, |b_i|)]^{l_i+u_i} \\
& \propto \frac{1}{[f(\omega)]^{l+u}} \frac{1}{\sigma^n} \prod_{i=1}^n f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \\
& \stackrel{c}{=} \frac{1}{[f(\omega)]^{l+u}} \frac{1}{\sigma^n} \prod_{i=1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{l_i+u_i} \\
& \stackrel{d}{=} \frac{\prod_{i=1}^p (1/\sigma)f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)}{\sigma^{k-p-1/2}} \left[ \frac{\omega/\sigma}{\omega f(\omega)} \right]^{l+u} \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{l_i+u_i}.
\end{aligned}$$

In step *a*, we use  $y_i = a_i + b_i\omega$  and Lemma 1 to obtain

$$f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) = f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma + a_i/\sigma) \leq D(|a_i|, 1)f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma),$$

because  $|a_i/\sigma| \leq |a_i|$  for all  $i$ . We also use  $\pi(\boldsymbol{\beta}, \sigma) \leq \max(\sigma^{-1}, 1)B = B$ . In step *b*, we use again Lemma 1 to obtain  $f(\omega)/f(y_i) = f((y_i - a_i)/b_i)/f(y_i) \leq |b_i|D(|a_i|, |b_i|)$ . In step *c*, we set  $b_i = 0$  if  $k_i = 1$  and we use the symmetry of  $f$  to obtain  $f(-\mathbf{x}_i^T \boldsymbol{\beta}/\sigma) = f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)$ . In step *d*, we use the assumption  $k_1 = \dots = k_p = 1$ .

Now it suffices to demonstrate that

$$\left[ \frac{\omega/\sigma}{\omega f(\omega)} \right]^{l+u} \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} [f((b_i\omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{l_i+u_i} \quad (4)$$

is bounded by a constant that does not depend on  $\omega, \boldsymbol{\beta}$  and  $\sigma$  since  $(1/\sigma)^{k-p-1/2} \prod_{i=1}^p (1/\sigma) \times f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)$  is an integrable function on area 1. Indeed, since  $k > p + 1$ , we have

$$\int_1^\infty (1/\sigma)^{k-p-1/2} \int_{\mathbb{R}^p} \prod_{i=1}^p (1/\sigma)f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma) d\boldsymbol{\beta} d\sigma = \left| \det \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_p \end{pmatrix} \right|^{-1} \int_1^\infty \frac{1}{\sigma^{k-p-1/2}} d\sigma < \infty,$$

using the following change of variables:  $u_i = \mathbf{x}_i^T \boldsymbol{\beta}/\sigma, i = 1, \dots, p$ . The determinant is different from 0 because  $\mathbf{x}_1, \dots, \mathbf{x}_p$  are linearly independent. Note that if instead, in step *a* above, we bound  $\pi(\boldsymbol{\beta}, \sigma)$  by  $\sigma^{-1}B$ , one can verify that the condition  $k \geq p + 1$  is sufficient to bound above the integral.

In order to bound above the function in (4), we separately analyse the three following cases:  $\omega/\sigma$  is large,  $\omega/\sigma$  is either large or bounded, and  $\omega/\sigma$  is bounded. More precisely, we split area 1 into three parts:  $1 \leq \sigma < \omega^{1/2}$ ,  $\omega^{1/2} \leq \sigma < \omega/(\zeta M)$  and  $\omega/(\zeta M) \leq \sigma < \infty$ , where  $M$  is defined in (3) and  $\zeta$  is a positive constant. Note that this split is well defined if  $y > \max(1, (\zeta M)^2)$  since  $\omega \geq y$ .

First, consider  $\omega/(\zeta M) \leq \sigma < \infty$ . We have

$$\begin{aligned} & \left[ \frac{\omega/\sigma}{\omega f(\omega)} \right]^{l+u} \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} \left[ f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{l_i+u_i} \stackrel{a}{\leq} \frac{B^{n-p}}{\sigma^{1/2}} \left[ \frac{\omega/\sigma}{\omega f(\omega)} \right]^{l+u} \\ & \stackrel{b}{\leq} B^{n-p} (\zeta M)^{l+u+1/2} \frac{(1/\omega)^{1/2}}{[\omega f(\omega)]^{l+u}} \stackrel{c}{\leq} B^{n-p} (\zeta M)^{l+u+1/2} \frac{(1/\omega)^{1/2}}{(\log \omega)^{-(\rho+1)(l+u)}} \\ & \stackrel{d}{\leq} B^{n-p} (\zeta M)^{l+u+1/2} [2(\rho+1)(l+u)/e]^{(\rho+1)(l+u)} < \infty. \end{aligned}$$

In step *a*, we use  $f \leq B$ . In step *b*, we use  $\omega/\sigma \leq \zeta M$  and  $1/\sigma \leq \zeta M/\omega$ . In step *c*, we use  $\omega f(\omega) > (\log \omega)^{-\rho-1}$  if  $\omega \geq y \geq A(1)$ , where  $A(1)$  comes from Proposition 2. For step *d*, it is purely algebraic to show that the maximum of  $(\log \omega)^\xi / \omega^{1/2}$  is  $(2\xi/e)^\xi$  for  $\omega > 1$  and  $\xi > 0$ , where  $\xi = (\rho+1)(l+u)$  in our situation.

Now, consider the two other parts combined (we will split them in the next step), that is  $1 \leq \sigma \leq \omega/(\zeta M)$ . We have,

$$\begin{aligned} & \left[ \frac{\omega/\sigma}{\omega f(\omega)} \right]^{l+u} \frac{1}{\sigma^{1/2}} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} \left[ f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{l_i+u_i} \\ & = \frac{1}{\sigma^{1/2}} \left[ \frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^{l+u} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} \left[ f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{l_i+u_i} \\ & \stackrel{a}{\leq} \frac{1}{\sigma^{1/2}} \left[ \frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^{l+u} B^{k-p} [D(0, \zeta)\zeta]^{l+u}. \end{aligned}$$

In step *a*, we use Lemma 2.

Now, we consider  $\omega^{1/2} \leq \sigma \leq \omega/(\zeta M)$ . We have

$$\begin{aligned} & \frac{1}{\sigma^{1/2}} \left[ \frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^{l+u} \stackrel{a}{\leq} B^{l+u} \frac{(1/\omega)^{1/4}}{[\omega f(\omega)]^{l+u}} \stackrel{b}{\leq} B^{l+u} \frac{(1/\omega)^{1/4}}{(\log \omega)^{-(\rho+1)(l+u)}} \\ & \stackrel{c}{\leq} B^{l+u} [4(\rho+1)(l+u)/e]^{(\rho+1)(l+u)} < \infty. \end{aligned}$$

In step *a*, we use  $(\omega/\sigma)f(\omega/\sigma) \leq B$  and  $(1/\sigma)^{1/2} \leq (1/\omega)^{1/4}$ . In step *b*, we use  $\omega f(\omega) > (\log \omega)^{-\rho-1}$  if  $\omega \geq y \geq A(1)$ , where  $A(1)$  comes from Proposition 2. In step *c*, it is purely algebraic to show that the maximum of  $(\log \omega)^\xi / \omega^{1/4}$  is  $(4\xi/e)^\xi$  for  $\omega > 1$  and  $\xi > 0$ , where  $\xi = (\rho+1)(l+u)$  in our situation.

Finally, we consider  $1 \leq \sigma \leq \omega^{1/2}$ . We have,

$$\frac{1}{\sigma^{1/2}} \left[ \frac{(\omega/\sigma)f(\omega/\sigma)}{\omega f(\omega)} \right]^{l+u} \stackrel{a}{\leq} \left[ \frac{\omega^{1/2} f(\omega^{1/2})}{\omega f(\omega)} \right]^{l+u} \stackrel{b}{\leq} 2^{(\rho+1)(l+u)} < \infty.$$

In step *a*, we use  $1/\sigma \leq 1$  and  $(\omega/\sigma)f(\omega/\sigma) \leq \omega^{1/2}f(\omega^{1/2})$  by the monotonicity of the tails of  $|z|f(z)$  since  $\omega/\sigma \geq \omega^{1/2} \geq y^{1/2} \geq M$  if  $y \geq M^2$ . In step *b*, we use  $\omega^{1/2}f(\omega^{1/2})/(\omega f(\omega)) \leq 2(1/2)^{-\rho} = 2^{\rho+1}$  if  $\omega \geq y \geq A(1, 2)$ , where  $A(1, 2)$  comes from the definition of log-regularly varying functions (see Definition 2).

**Area 2:** Consider  $0 < \sigma < 1$ . We actually need to show that

$$\begin{aligned} \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} d\sigma d\boldsymbol{\beta} \\ = \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta}. \end{aligned}$$

For area 2, we proceed in a slightly different manner than for area 1. We begin by separating the first integral above into two parts as follows:

$$\begin{aligned} \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} d\sigma d\boldsymbol{\beta} \\ = \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \mathbb{1}_{\cap_j \mathcal{O}_j^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\ + \lim_{\omega \rightarrow \infty} \int_{\cup_j \mathcal{O}_j} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} d\sigma d\boldsymbol{\beta}, \end{aligned}$$

where

$$\mathcal{O}_j := \{\boldsymbol{\beta} : |y_j - \mathbf{x}_j^T \boldsymbol{\beta}| < \omega/2\}, \forall j \in \mathcal{I}_O,$$

with  $\mathcal{I}_O := \{i : i \in \{1, \dots, n\} \text{ and } l_i + u_i = 1\}$ . We show that the first part is equal to the integral  $\int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta}$  and that the second part is equal to 0.

For the first part, we again use Lebesgue's dominated convergence theorem in order to interchange the limit  $\omega \rightarrow \infty$  and the integral. We have



$$\begin{aligned}
& \lim_{\omega \rightarrow \infty} \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \mathbb{1}_{\cap_j \mathcal{O}_j^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \lim_{\omega \rightarrow \infty} \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \mathbb{1}_{\cap_j \mathcal{O}_j^c}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \times 1 \times \mathbb{1}_{\mathbb{R}^p}(\boldsymbol{\beta}) d\sigma d\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\sigma d\boldsymbol{\beta},
\end{aligned}$$

using Proposition 3 in the second equality since  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are fixed, and  $\lim_{\omega \rightarrow \infty} \mathbb{1}_{\cap_j \mathcal{O}_j^c}(\boldsymbol{\beta}) = \mathbb{1}_{\mathbb{R}^p}(\boldsymbol{\beta})$ . Indeed, if  $u_j = 1$ ,  $|y_j - \mathbf{x}_j^T \boldsymbol{\beta}| < \omega/2 \leq y_j/2 \Leftrightarrow y_j/2 < \mathbf{x}_j^T \boldsymbol{\beta} < 3y_j/2$ , and in the limit, no  $\boldsymbol{\beta} \in \mathbb{R}^p$  satisfies this (we have the same conclusion if  $l_j = 1$ ). Note that pointwise convergence is sufficient, for any value of  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\sigma > 0$ , once the limit is inside the integral. We now demonstrate that the integrand is bounded, for any value of  $\omega \geq y$ , by an integrable function of  $\boldsymbol{\beta}$  and  $\sigma$  that does not depend on  $\omega$ .

Consider  $\boldsymbol{\beta} \in \cap_j \mathcal{O}_j^c$ , that is  $\{\boldsymbol{\beta} : |y_j - \mathbf{x}_j^T \boldsymbol{\beta}| \geq |y_j|/2 \text{ for all } j \in \mathcal{I}_O\}$ , and  $0 < \sigma < 1$ . Note that the integrand is equal to 0 if  $\boldsymbol{\beta} \notin \cap_j \mathcal{O}_j^c$ . For all  $j \in \mathcal{I}_O$ , we have

$$(1/\sigma)f((y_j - \mathbf{x}_j^T \boldsymbol{\beta})/\sigma) \leq f(y_j - \mathbf{x}_j^T \boldsymbol{\beta}) \leq f(y_j/2) \leq 2D(0, 2)f(y_j),$$

by the monotonicity of the tails of  $|z|f(z)$  and then the monotonicity of the tails of  $f(z)$ , because  $|y_j - \mathbf{x}_j^T \boldsymbol{\beta}|/\sigma \geq |y_j - \mathbf{x}_j^T \boldsymbol{\beta}| \geq \omega/2 \geq y/2 \geq M$ , if we choose  $y \geq 2M$ . Lemma 1 is used in the last inequality. Therefore,

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \mathbb{1}_{\cap_j \mathcal{O}_j^c}(\boldsymbol{\beta}) \stackrel{a}{\leq} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) [2D(0, 2)]^{l+u},$$

which is an integrable function.

We now prove that

$$\lim_{\omega \rightarrow \infty} \int_{\cup_j \mathcal{O}_j} \int_0^1 \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} d\sigma d\boldsymbol{\beta} = 0.$$

We first bound above the integrand and then we prove that the integral of the upper bound converges towards 0 as  $\omega \rightarrow \infty$ . For the rest of the proof, we assume without loss of generality that  $y_1, \dots, y_{l+u+2p-1}$  are  $l+u+2p-1$  nonoutliers (therefore  $k_1 = \dots = k_{l+u+2p-1} = 1$ ) such that any  $p$  vectors picked among  $\mathbf{x}_1, \dots, \mathbf{x}_{l+u+2p-1}$  are linearly independent. In the same manner as in the proof of Lemma 2, we bound above the function on  $p+1$  mutually exclusive areas using

$$\begin{aligned} \cup_i \mathcal{O}_i = & \left[ \cup_j \left( \mathcal{O}_j \cap \left( \cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[ \cup_{j, i_1} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \left( \cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right] \\ & \cup \dots \cup \left[ \cup_{j, i_1, \dots, i_{p-1}} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left( \cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right] \\ & \cup \left[ \cup_{j, i_1, \dots, i_p} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} \right) \right], \end{aligned}$$

where

$$\mathcal{F}_i := \{\boldsymbol{\beta} : |\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\zeta\}, \forall i \in \mathcal{I}_{\mathcal{F}},$$

and  $\mathcal{I}_{\mathcal{F}} := \{1, \dots, l+u+2p-1\}$ . As explained in the proof of Lemma 2,  $\mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} = \emptyset$  for all  $j, i_1, \dots, i_p$ .

Now, we consider that  $0 < \sigma < 1$  and that  $\boldsymbol{\beta}$  belongs to one of the sets  $\mathcal{O}_j \cap \left( \cap_{i_1} \mathcal{F}_{i_1}^c \right)$ ,  $\left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \left( \cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right)$ ,  $\dots$ , or  $\left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left( \cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right)$ , and we bound the function. We have

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) & \prod_{i=1}^n \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i+u_i} \\ & \stackrel{a}{\leq} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{|b_i| D(|a_i|, |b_i|) (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{l_i+u_i} \\ & \propto \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[ (1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{l_i+u_i} \\ & \stackrel{b}{\leq} (B/\sigma) [2\zeta D(0, 2\zeta) (1/\sigma) f(\omega/\sigma)]^{l+u+1} \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n \left[ (1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \\ & \quad \times \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{l_i+u_i} \\ & \propto (1/\sigma) [(1/\sigma) f(\omega/\sigma)]^{l+u+1} \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n \left[ (1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \\ & \quad \times \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega)} \right]^{l_i+u_i} \\ & \stackrel{c}{\leq} (1/\sigma) (1/\sigma) f(\omega/\sigma) \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n \left[ (1/\sigma) f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i} \left[ (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{l_i+u_i} \\ & \stackrel{d}{\leq} B\omega^{-1} (1/\sigma) \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma). \end{aligned}$$

In step *a*, we use Lemma 1 to obtain  $f(\omega)/f(y_i) = f((y_i - a_i)/b_i)/f(y_i) \leq |b_i|D(|a_i|, |b_i|)$  for all  $i \in \mathcal{I}_O$ . In step *b*, we use  $\pi(\boldsymbol{\beta}, \sigma) \leq \max(\sigma^{-1}, 1)B = B/\sigma$  and the fact that in any of the sets in which  $\boldsymbol{\beta}$  can belong, there are at least  $l + u + p$  nonoutlying points  $(\mathbf{x}_i, a_i)$  such that  $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega/\zeta$  (note that  $l + u + p \geq l + u + 2$  because we only consider the models with  $p \geq 2$ ). This implies that there exists  $\{i_1, \dots, i_{l+u+1}\} \subset \mathcal{I}_{\mathcal{F}}$  such that for all  $i \in \{i_1, \dots, i_{l+u+1}\}$ ,

$$f((a_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \leq f(\omega/(2\zeta\sigma)) \leq 2\zeta D(0, 2\zeta) f(\omega/\sigma),$$

using the monotonicity of the tails of  $f$  in the first inequality because, if we define the constant  $a_{(k)} := \max_{i \in \{1, \dots, k\}} |a_i|$  with  $\omega \geq y \geq (2\zeta)a_{(k)}$ , we have  $|a_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma \geq (|\mathbf{x}_i^T \boldsymbol{\beta}| - |a_i|)/\sigma \geq (\omega/\zeta - a_{(k)})/\sigma \geq \omega/(2\zeta\sigma) \geq \omega/(2\zeta) \geq y/(2\zeta) \geq M$  if we choose  $y \geq 2\zeta M$ . We use Lemma 1 in the second inequality (as mentioned in the proof of Lemma 2, we choose  $\zeta \geq 1$ ). In step *c*, we use the monotonicity of the tails of  $|z|f(z)$  to obtain  $(\omega/\sigma)f(\omega/\sigma) \leq \omega f(\omega)$  for  $l + u$  terms, because  $\omega/\sigma \geq \omega \geq y \geq M$  if we choose  $y \geq M$ . In step *d*, we use  $(1/\sigma)f(\omega/\sigma) \leq B/\omega$ .

The integral of  $B\omega^{-1}(1/\sigma) \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$  is bounded by

$$B\omega^{-1} \int_{\mathbb{R}^p} \int_0^\infty (1/\sigma) \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\sigma d\boldsymbol{\beta} = B\omega^{-1} m(y_{i_{l+u+2}}, \dots, y_{i_n}),$$

where  $m(y_{i_{l+u+2}}, \dots, y_{i_n})$  is the marginal density arising from a prior proportional to  $1/\sigma$  and  $n - (l + u + 1) = k - 1$  observations  $(\mathbf{x}_{i_j}, y_{i_j})$ ,  $j = l + u + 2, \dots, n$ ,  $\{i_{l+u+2}, \dots, i_n\}$  being the set  $\{1, \dots, n\} \setminus \{i_1, \dots, i_{l+u+1}\}$ . In order to prove that  $B\omega^{-1} m(y_{i_{l+u+2}}, \dots, y_{i_n}) \rightarrow 0$  as  $\omega \rightarrow \infty$ , it suffices to prove that  $m(y_{i_{l+u+2}}, \dots, y_{i_n})$  is bounded above by a constant that does not depend on  $\omega$ , because  $\omega^{-1} \rightarrow 0$ . In the proof of Proposition 1, we proved that  $m(y_{i_{l+u+2}}, \dots, y_{i_n})$  is bounded above by a constant that does not depend on  $\omega$  if the number of observations is greater than or equal to  $p + 1$  (as mentioned in this proof, the condition requiring that the number of observations be greater than  $p + 1$  can be relaxed when the prior is proportional to  $1/\sigma$ ), and if among  $\mathbf{x}_{i_{l+u+2}}, \dots, \mathbf{x}_{i_n}$  there are  $p$  linearly independent vectors. Since we assumed that  $l + u \geq 1$  (the proof for the case  $l + u = 0$  is trivial),  $k \geq l + u + 2p - 1$ , and that  $y_1, \dots, y_{l+u+2p-1}$  are  $l + u + 2p - 1$  nonoutliers such that any  $p$  vectors picked among  $\mathbf{x}_1, \dots, \mathbf{x}_{l+u+2p-1}$  are linearly independent, it implies that  $m(y_{i_{l+u+2}}, \dots, y_{i_n})$  is the marginal of  $k - 1 \geq p + 1$  observations ( $p \geq 2$ ), where  $p$  of them have linearly independent  $\mathbf{x}$  vectors. As a result,

$$B\omega^{-1} \int_{\mathbb{R}^p} \int_0^\infty (1/\sigma) \prod_{i=1(i \neq i_1, \dots, i_{l+u+1})}^n (1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) d\sigma d\boldsymbol{\beta} \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

We therefore have that

$$\int_{\cup_i \mathcal{O}_i} \int_0^1 \pi(\boldsymbol{\beta}, \sigma | \mathbf{y}_k) \prod_{i=1}^n \left[ \frac{(1/\sigma)f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i + u_i} d\sigma d\boldsymbol{\beta} \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

□

*Proof of Result (b).* Consider  $(\boldsymbol{\beta}, \sigma)$  such that  $\pi(\boldsymbol{\beta}, \sigma) > 0$  (the proof for the case  $(\boldsymbol{\beta}, \sigma)$  such that  $\pi(\boldsymbol{\beta}, \sigma) = 0$  is trivial). We have, as  $\omega \rightarrow \infty$ ,

$$\begin{aligned} \frac{\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)}{\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)} &= \frac{m(\mathbf{y}_k)}{m(\mathbf{y}_n)} \times \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[ (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{k_i}} \\ &= \frac{m(\mathbf{y}_k)}{m(\mathbf{y}_n)} \prod_{i=1}^n \left[ (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right]^{l_i + u_i} \\ &= \frac{m(\mathbf{y}_k) \prod_{i=1}^n [f(y_i)]^{l_i + u_i}}{m(\mathbf{y}_n)} \prod_{i=1}^n \left[ \frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)} \right]^{l_i + u_i} \rightarrow 1. \end{aligned}$$

The first ratio in the last equality does not depend on  $\boldsymbol{\beta}$  and  $\sigma$  and converges towards 1 as  $\omega \rightarrow \infty$  using result (a). The product also converges towards 1 uniformly in any set  $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$  using Proposition 3 since  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are fixed. Furthermore, since  $f$  and  $\min(\sigma, 1)\pi(\boldsymbol{\beta}, \sigma)$  are bounded,  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  is also bounded on any set  $(\boldsymbol{\beta}, \sigma) \in [-\lambda, \lambda]^p \times [1/\tau, \tau]$ . Then, we have

$$\left| \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) - \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \right| = \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) \left| \frac{\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)}{\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)} - 1 \right| \rightarrow 0 \text{ as } \omega \rightarrow \infty.$$

□

*Proof of Results (c) and (d).* Using Proposition 1, we know that  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  and  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n)$  are proper. Moreover, using result (b), we have the pointwise convergence  $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) \rightarrow \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k)$  as  $\omega \rightarrow \infty$  for any  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\sigma > 0$ , as a result of the uniform convergence. Then, the conditions of Scheffé's theorem (see Scheffé (1947)) are satisfied and we obtain the convergence in  $L_1$  given by result (c) as well as the following result:

$$\lim_{\omega \rightarrow \infty} \int_E \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_n) d\boldsymbol{\beta} d\sigma = \int_E \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y}_k) d\boldsymbol{\beta} d\sigma,$$

uniformly for all sets  $E \subset \mathbb{R}^p \times \mathbb{R}^+$ . Result (d) follows directly. □

*Proof of Result (e).* Using equation (1), result (e) follows directly from result (b). □

We now present two propositions and two lemmas that are used in the proof of Theorem 1. The proofs of Propositions 2 and 3, and of Lemma 1 can be found in Desgagné (2015).

**Proposition 2** (Dominance). *If  $s \in L_0(\infty)$  and  $g \in L_\rho(\infty)$ , then for all  $\delta > 0$ , there exists a constant  $A(\delta) > 1$  such that  $z \geq A(\delta) \Rightarrow$*

$$(\log z)^{-\delta} < s(z) < (\log z)^\delta \quad \text{and} \quad (\log z)^{-\rho-\delta} < g(z) < (\log z)^{-\rho+\delta}.$$

**Proposition 3** (Location-scale transformation). *If  $zf(z) \in L_\rho(\infty)$ , then we have*

$$(1/\sigma)f((z - \mu)/\sigma)/f(z) \rightarrow 1 \text{ as } z \rightarrow \infty,$$

*uniformly on  $(\mu, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau]$ , for any  $\lambda \geq 0$  and  $\tau \geq 1$ .*

**Lemma 1.**  $\forall \lambda \geq 0, \forall \tau \geq 1$ , there exists a constant  $D(\lambda, \tau) \geq 1$  such that  $z \in \mathbb{R}$  and  $(\mu, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau] \Rightarrow$

$$1/D(\lambda, \tau) \leq (1/\sigma)f((z - \mu)/\sigma)/f(z) \leq D(\lambda, \tau).$$

Note that Lemma 1 is a corollary of Proposition 3.

**Lemma 2.** Consider that  $y_1, \dots, y_p$  are  $p$  nonoutliers (therefore  $k_1 = \dots = k_p = 1$ ) such that  $\mathbf{x}_1, \dots, \mathbf{x}_p$  are linearly independent. If  $1 \leq \sigma \leq \omega/(\zeta M)$  and  $\boldsymbol{\beta} \in \mathbb{R}^p$ , then

$$\prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta}/\sigma)]^{k_i} \left[ \frac{f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(\omega/\sigma)} \right]^{l_i+u_i} \leq B^{k-p} [D(0, \zeta) \zeta]^{l+u},$$

where  $M$  is defined in (3) and  $\zeta$  is a positive constant.

*Proof of Lemma 2.* In Theorem 1, we assume that there exists  $(y_{i_1}, \dots, y_{i_{l+u+2p-1}}) \subset \mathbf{y}_k$  such that any  $p$  vectors picked among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{l+u+2p-1}}$  are linearly independent. We know that  $y_1, \dots, y_p$  are  $p$  nonoutliers (therefore  $k_1 = \dots = k_p = 1$ ) such that  $\mathbf{x}_1, \dots, \mathbf{x}_p$  are linearly independent. For this proof, we assume without loss of generality that  $y_{p+1}, \dots, y_{l+u+2p-1}$  are  $l+u+p-1$  nonoutliers (therefore  $k_{p+1} = \dots = k_{l+u+2p-1} = 1$ ) such that any  $p$  vectors picked among  $\mathbf{x}_{p+1}, \dots, \mathbf{x}_{l+u+2p-1}$  are linearly independent. Note that there are at least  $p$  of these nonoutliers (because we assume that  $l+u \geq 1$ ).

In order to prove the result, we split the domain of  $\boldsymbol{\beta}$  into  $p+2$  mutually exclusive areas as follows:

$$\begin{aligned} \mathbb{R}^p = & \left[ \cap_j \mathcal{O}_j^c \right] \cup \left[ \cup_j \left( \mathcal{O}_j \cap \left( \cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[ \cup_{j, i_1} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \left( \cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right] \\ & \cup \dots \cup \left[ \cup_{j, i_1, \dots, i_{p-1}} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left( \cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right] \\ & \cup \left[ \cup_{j, i_1, \dots, i_p} \left( \mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} \right) \right], \end{aligned}$$

where

$$\begin{aligned} \mathcal{O}_j &:= \{\boldsymbol{\beta} : |b_j \omega - \mathbf{x}_j^T \boldsymbol{\beta}| < \omega/2\}, \forall j \in \mathcal{I}_O, \\ \mathcal{F}_i &:= \{\boldsymbol{\beta} : |\mathbf{x}_i^T \boldsymbol{\beta}| < \omega/\zeta\}, \forall i \in \mathcal{I}_{\mathcal{F}}, \end{aligned}$$

$\mathcal{I}_O := \{i : i \in \{p+1, \dots, n\} \text{ and } l_i + u_i = 1\}$  and  $\mathcal{I}_{\mathcal{F}} := \{p+1, \dots, l+u+2p-1\}$  are the sets of indexes of outliers and remaining fixed observations (nonoutliers) that are such that any  $p$  vectors picked among  $\mathbf{x}_{p+1}, \dots, \mathbf{x}_{l+u+2p-1}$  are linearly independent, respectively.

To fix ideas, the set  $\mathcal{O}_j$  represents hyperplanes passing relatively near (at a vertical distance of  $\omega/2$  or less) the point  $(\mathbf{x}_j, b_j \omega)$ , which is considered as an outlier since  $\omega \rightarrow \infty$ . The set  $\mathcal{F}_i$  represents hyperplanes passing relatively near (at a vertical distance of  $\omega/\zeta$  or less) the point  $(\mathbf{x}_i, 0)$ , which is considered as a nonoutlier. Therefore, the set  $\cap_j \mathcal{O}_j^c$  represents hyperplanes passing

not “near” any outliers. For each  $j \in \mathcal{I}_O$ , the set  $O_j \cap (\cap_{i_1} \mathcal{F}_{i_1}^c)$  represents hyperplanes passing “near” the point  $(\mathbf{x}_j, b_j\omega)$  (an outlier) and possibly near other outliers, but not “near” any nonoutliers. For each  $j \in \mathcal{I}_O$  and  $i_1 \in \mathcal{I}_\mathcal{F}$ , the set  $O_j \cap \mathcal{F}_{i_1} \cap (\cap_{i_2} \mathcal{F}_{i_2}^c)$  represents hyperplanes passing “near” the point  $(\mathbf{x}_j, b_j\omega)$  (an outlier) and possibly near other outliers, and the point  $(\mathbf{x}_{i_1}, 0)$  (a nonoutlier), but not “near” other nonoutliers. And so on.

We first show that the function is bounded on  $\beta \in \cap_j O_j^c$  and  $1 \leq \sigma \leq \omega/(\zeta M)$ . For all  $j \in \mathcal{I}_O$ , we have

$$\frac{f((b_j\omega - \mathbf{x}_j^T \beta)/\sigma)}{f(\omega/\sigma)} \leq \frac{f(\omega/(2\sigma))}{f(\omega/\sigma)} \leq 2D(0, 2) \leq D(0, \zeta)\zeta,$$

using the monotonicity of  $f$  because  $|b_j\omega - \mathbf{x}_j^T \beta|/\sigma \geq \omega/(2\sigma) \geq \zeta M/2 \geq M$  (we choose  $\zeta \geq 2$ ), and then Lemma 1. Therefore,

$$\prod_{i=p+1}^n [f(\mathbf{x}_i^T \beta/\sigma)]^{k_i} \left[ \frac{f((b_i\omega - \mathbf{x}_i^T \beta)/\sigma)}{f(\omega/\sigma)} \right]^{l_i+u_i} \leq B^{k-p} [D(0, \zeta)\zeta]^{l+u}.$$

Now, we claim that  $O_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} = \emptyset$  if  $\mathbf{x}_j, \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$  are linearly dependent. To prove this, we consider the two following distinct cases:  $\mathbf{x}_j$  is a linear combination of  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$ , and one vector among  $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$ , say  $\mathbf{x}_{i_1}$ , is a linear combination of  $\mathbf{x}_j$  and  $\mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_{p-1}}$ . First, considering that  $\beta \in \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}}$  and  $\mathbf{x}_j = \sum_{s=1}^{p-1} a_s \mathbf{x}_{i_s}$ ,  $a_1, \dots, a_{p-1} \in \mathbb{R}$ ,  $|b_j\omega - \mathbf{x}_j^T \beta| = |b_j\omega - (\sum_{s=1}^{p-1} a_s \mathbf{x}_{i_s})^T \beta| \geq |b_j\omega| - |(\sum_{s=1}^{p-1} a_s \mathbf{x}_{i_s})^T \beta| \geq \omega - \sum_{s=1}^{p-1} |a_s| \omega/\zeta$ , which is greater than or equal to  $\omega/2$  if  $\zeta \geq 2 \sum_{s=1}^{p-1} |a_s|$  (we choose  $\zeta$  such that it satisfies this inequality for any combination of  $j$  and  $i_1, \dots, i_{p-1}$ ). Therefore, we have that  $\beta \notin O_j$ . Second, considering that  $\beta \in O_j \cap \mathcal{F}_{i_2} \cap \dots \cap \mathcal{F}_{i_{p-1}}$  and  $\mathbf{x}_{i_1} = a_1 \mathbf{x}_j + \sum_{s=2}^{p-1} a_s \mathbf{x}_{i_s}$ ,  $a_1, \dots, a_{p-1} \in \mathbb{R}$ ,  $|\mathbf{x}_{i_1}^T \beta| = |(a_1 \mathbf{x}_j + \sum_{s=2}^{p-1} a_s \mathbf{x}_{i_s})^T \beta| \geq (|a_1|/2)|b_j\omega| - |(\sum_{s=2}^{p-1} a_s \mathbf{x}_{i_s})^T \beta| \geq |a_1| \omega/2 - \sum_{s=2}^{p-1} |a_s| \omega/\zeta$ , which is greater than or equal to  $|a_1| \omega/4$  if  $\zeta \geq 4 \sum_{s=2}^{p-1} |a_s|/|a_1|$  (we choose  $\zeta$  such that it satisfies this inequality for any combination of  $j$  and  $i_1, \dots, i_{p-1}$ ). Therefore, we have that  $\beta \notin \mathcal{F}_{i_1}$  (we choose  $\zeta > 4/|a_1|$ ). Note that  $a_1 \neq 0$ , because  $a_1 = 0$  means that  $\mathbf{x}_{i_1}$  is a linear combination of  $\mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_{p-1}}$ , but we know that  $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_{p-1}}$  are linearly independent. Note also that we use the fact that  $|\mathbf{x}_j^T \beta| \geq |b_j| \omega/2$ . We explain this now: if  $u_j = 1$ ,  $|b_j\omega - \mathbf{x}_j^T \beta| < \omega/2 \leq |b_j| \omega/2$ , which is equivalent to  $-b_j\omega/2 < b_j\omega - \mathbf{x}_j^T \beta < b_j\omega/2 \Rightarrow |\mathbf{x}_j^T \beta| \geq |b_j| \omega/2$  (and we have the same inequality if  $l_j = 1$ ). All this proves that  $O_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} = \emptyset$  if  $\mathbf{x}_j, \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_{p-1}}$  are linearly dependent.

Using the same arguments as above, we know that  $O_j \cap \mathcal{F}_{i_1} = \emptyset$  if  $\mathbf{x}_j, \mathbf{x}_{i_1}$  are linearly dependent,  $O_j \cap \mathcal{F}_{i_1} \cap \mathcal{F}_{i_2} = \emptyset$  if  $\mathbf{x}_j, \mathbf{x}_{i_1}, \mathbf{x}_{i_2}$  are linearly dependent, and so on. In particular, we know that  $O_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_p} = \emptyset$  for all  $j, i_1, \dots, i_p$  because we are sure that  $\mathbf{x}_j, \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_p}$  are linearly dependent (because of the dimension of the space). This implies that

$$\begin{aligned} \mathbb{R}^p = & \left[ \cap_j O_j^c \right] \cup \left[ \cup_j \left( O_j \cap \left( \cap_{i_1} \mathcal{F}_{i_1}^c \right) \right) \right] \cup \left[ \cup_{j, i_1} \left( O_j \cap \mathcal{F}_{i_1} \cap \left( \cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c \right) \right) \right] \\ & \cup \dots \cup \left[ \cup_{j, i_1, \dots, i_{p-1}} \left( O_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap \left( \cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c \right) \right) \right]. \end{aligned}$$

Now, we consider that  $1 \leq \sigma \leq \omega/(\zeta M)$  and that  $\boldsymbol{\beta}$  belongs to one of sets  $\mathcal{O}_j \cap (\cap_{i_1} \mathcal{F}_{i_1}^c)$ ,  $(\mathcal{O}_j \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))$ ,  $\dots$ , or  $(\mathcal{O}_j \cap \mathcal{F}_{i_1} \cap \dots \cap \mathcal{F}_{i_{p-1}} \cap (\cap_{i_p \neq i_1, \dots, i_{p-1}} \mathcal{F}_{i_p}^c))$ , and we bound the function. We have

$$\begin{aligned} \prod_{i=p+1}^n [f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i} \left[ \frac{f((b_i \omega - \mathbf{x}_i^T \boldsymbol{\beta}) / \sigma)}{f(\omega / \sigma)} \right]^{l_i + u_i} &\stackrel{a}{\leq} B^{l+u} \prod_{i=p+1}^n \frac{[f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma)]^{k_i}}{f(\omega / \sigma)^{l_i + u_i}} \\ &\stackrel{b}{\leq} B^{l+u+(k-p)-(l+u)} [D(0, \zeta) \zeta]^{l+u} = B^{k-p} [D(0, \zeta) \zeta]^{l+u}. \end{aligned}$$

In step *a*, we use  $f \leq B$  for all  $i \in \mathcal{I}_\mathcal{O}$ . In step *b*, we use the fact that in any of the sets in which  $\boldsymbol{\beta}$  can belong, there are at least  $l + u$  nonoutlying points  $(\mathbf{x}_i, 0)$  such that  $|\mathbf{x}_i^T \boldsymbol{\beta}| \geq \omega / \zeta$ . This implies that

$$f(\mathbf{x}_i^T \boldsymbol{\beta} / \sigma) / f(\omega / \sigma) \leq f(\omega / (\zeta \sigma)) / f(\omega / \sigma) \leq D(0, \zeta) \zeta, \quad (5)$$

using the monotonicity of  $f$  because  $|\mathbf{x}_i^T \boldsymbol{\beta}| / \sigma \geq \omega / (\zeta \sigma) \geq M$ , and then Lemma 1. For the remaining  $k - p - (l + u)$  nonoutlying points, we use  $f \leq B$ . Note that we need to have  $k - p - (p - 1) \geq l + u$ . □

## References

## References

- Andrade, J. A. A., O'Hagan, A., 2011. Bayesian robustness modelling of location and scale parameters. *Scand. J. Stat.* 38 (4), 691–711.
- Box, G. E. P., Tiao, G. C., 1968. A bayesian approach to some outlier problems. *Biometrika* 55 (1), 119–129.
- Desgagné, A., 2013. Full robustness in bayesian modelling of a scale parameter. *Bayesian Anal.* 8 (1), 187–220.
- Desgagné, A., 2015. Robustness to outliers in location–scale parameter model using log-regularly varying distributions. *Ann. Statist.* 43 (4), 1568–1595.
- Desgagné, A., Gagnon, P., 2016. Bayesian robustness to outliers in linear regression and ratio estimation, article submitted for publication.
- Kass, R. E., Raftery, A. E., 1995. Bayes factors. *J. Amer. Statist. Assoc.* 90 (430), 773–795.
- O'Hagan, A., Pericchi, L., 2012. Bayesian heavy-tailed models and conflict resolution: A review. *Braz. J. Probab. Stat.* 26 (4), 372–401.
- Peña, D., Zamar, R., Yan, G., 2009. Bayesian likelihood robustness in linear models. *J. Statist. Plann. Inference* 139 (7), 2196 – 2207.
- Scheffé, H., 1947. A useful convergence theorem for probability distributions. *Ann. Math. Statist.*, 434–438.
- West, M., 1984. Outlier models and prior distributions in bayesian linear regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 46 (3), 431–439.